Comp4801 Final Year Project

AI-Driven Blockchain Forensics

Detailed Project Plan

Group: FYP23012

Supervised by:
Dr. Chow, Kam Pui

Member:
Chow Pak Hang (3035787920) – BEng(CompSc)

Content

1 Background

The emergence of blockchain technology have significantly influencing the development of financial technology (FinTech) worldwide, which provide a more secure environment for transaction, in addition, allowing efficient payment without approval of single centralized party [1]. While the cryptocurrency established on blockchain, such as Bitcoin and Ethereum have increasing popularity in the field of investment and transaction [1]. On the other hand, blockchain technology is defined as the trust machine for transaction based on mathematical calculation [1].

Although blockchain technology could be beneficial in financial activities and capital market as contrast to traditional financial system, criminal activities on blockchain are being more prevalent in terms of scams, phishing, money-laundering, and underground market activities, which could threaten the security of transaction and hinder the further development of blockchain in financial area [1, 2]. It is reported that there is more than 10 billion USD loss through illegal transaction in blockchain [3], indicating the substantial impacts to the society. Due to immutable property of blockchain, all transactions are irreversible, and limited opportunity to get refunds that involved of cryptocurrency-related criminal activities [2]. Furthermore, the decentralized property of blockchain have increase the difficulties in tracking and forensics of these transaction, in particular, most of the cryptocurrency have provided pseudo-anonymous identifiers that hide the identity of the sender and recipients [4]. Therefore, the investigation of illegal activities within blockchain is becoming crucial recently.

While it is true that a variety of blockchain explorer websites are publicly available to search for information about the specific account and transaction address without any technical knowledge. Nonetheless, research and applications on the blockchain data analysis and risk detection are limited and diverse respectively, for instance, analysis of Bitcoin account cannot be applied to Ethereum platform due to the differences in their account models [4, 5]. Hence, the motivation for analyzing blockchain data with Artificial Intelligence (AI) is the risks associated with the malicious transaction on blockchain. In the investigation, this project would be proposed for leveraging machine learning algorithms in detection of illicit transaction and analysis blockchain data in certain depth.

2 Related Work

2.1 Common Algorithms in Anomalies Detection

In classifying among normal and anomaly account, certain research has shown the utilization of various machine learning algorithms.

As in [6], **XGBoost** is employed in recognizing anomalies within Ethereum platform. **XGBoost** is a novel machine learning algorithms based on decision tree and gradient boosting techniques, it could provide efficient and reliable performance in presence of limited data among existing ensemble approaches [6-7].

The recognition of illicit account may involve of the transaction network in the blockchain, this could be interpreted as graph network for further investigation. Wu et al. [8] proposed the integration of transaction data into graph-based structure through a **Network Embedding algorithm**. Additionally, the generated node embeddings are processed through one-class **Support Vector Machine (SVM)** method for detecting phishing accounts [8]. It is indicated that more accurate and efficient feature extraction could be conducted via the network embedding, thus, the effectiveness of classification could be enhanced [8].

Moreover, there are several approaches in utilizing **Random Forest (RF)**, **Logistic Regression (LR)**, **Multilayer Perceptrons (MLP)**, and **Graph Convolutional Networks (GCN)** for recognition of illegal transaction in Bitcoin and Ethereum blockchain [9,10].

2.2 Limitation in Anomalies Detection

Most of the strategies [6-10] would be based on supervised learning, which may pose potential limitations to effectiveness of the classifier model. Since labelling for dataset is necessary in supervised learning, accuracy of these labels would definitely affect the performance of the machine learning algorithms. However, there is lacking verified labelling, in addition, the amount of illicit account is inadequate as compared to the substantial amount of normal account [5, 11]. Consequently, the quality of validation in classifying anomaly account may be suffered [5]. Despite there are several proposed dataset from different research, various obstacles such as low accuracy, ambiguity of features, and insufficient dataset are encountered [11].

3 Objective

In this project, the main objectives would address in three aspects, improving traceability, detection of malicious activities within blockchain, and compare and evaluate existing models on blockchain forensics.

Objective 1: Improvement of Traceability

One of the objective is to improve the traceability of account within blockchain. Since transactions of blockchain data are complex and considerable in amount, tracking transactions with trivial blockchain explorer would be inefficient and time-consuming. For instance, it may be difficult to trace the transaction flow by searching manually. Hence, this project proposed to collect and analyze real-time blockchain data in a more recognized way, through a graphical visualization approaches, which allow user to identify the possible address and account involved of the specific account or transaction.

Objective 2: Detection of Malicious activities

Detecting certain illegal activities would be another focus of the project. In public blockchain, the absence of labelling of different account could be unfavorable for the public to recognize criminal activities such as fraud on the blockchain. With the help of machine learning algorithm and data collected, this project could identify potential risks and illicit activities for the specific account and transaction. Thus, this could protect user from scams and frauds. The model would be mainly involved of classification algorithms, which distinguish between normal and illicit accounts, moreover, analyze and differentiate patterns of illegal activity that could be valuable for further research.

Objective 3: Evaluation of Existing Model

The project is proposed to compare and evaluate various machine learning for blockchain forensics in recognizing criminal activities. Recently, there are platforms such as Chainalysis and Elliptic to monitor and investigate within different blockchains. Nevertheless, their approaches and strategies for model training may not be visible to investigator directly [4]. Furthermore, existing research [6, 12-14] on crime detection in blockchain have used various machine learning algorithms and data representations. Therefore, it is believed that the evaluation of different algorithms would provide insight towards blockchain forensics.

4 Scope

In this project, it would primarily target on the Ethereum Blockchain among various blockchain.

In public (or permissionless) blockchain, all the transaction data are accessible and immutable on the blockchain, this allows more accurate and effective data collection. While illegal activities would be more prevalent in Bitcoin (the origination of blockchain in 2009), and it has the most widely usage of blockchain [14]. The newer blockchain platform. Ethereum, founded in 2015, is growing rapidly, and allow more usage in terms of smart contracts.

In Ethereum blockchain platform, the user operations are maintained in the environment of Ethereum Virtual Machine (EVM). In Ethereum, there are two types of accounts: External Owned Account (EOA) and Contract Owned Account (COA). For CA, it cannot initiate transaction alone while all transactions are initiated by EOA [15]. Since analysis of smart contract code in COA would not be the focus of this project, thus, the project would rather point toward the transaction of EOA.

On the other hand, illicit accounts would be generally referring transactions related to phishing, scams, Ponzi schemes, honeypot attacks, money laundering, and other illegal activities.

5 Methodology

To implement blockchain forensics application, the project would be divided into three stages generally. In the first stage, the development would be handling classification of illicit account or activities by leveraging certain machine learning algorithms (Section 5.1). In the second stage, it would be focus on organizing and analyzing of the blockchain data in specific account (Section 5.2). The final stage would be the front-end development and integration with deliverables from previous stags (Section 5.3).

5.1 Anomaly Account Recognition

5.1.1 Data Preparation

In detecting malicious account (or transaction) in blockchain, it may rely on the training of machine learning classification model with pre-defined dataset. Since there is no attribute of account labelling on public blockchain, several blockchain explorers and Open-Source Intelligence (OSINT) would be utilized for labelled dataset (i.e., labelling for account related to criminal activities such as scams, phishing, and fraud). For instance, combination of labelled accounts and datasets from Etherscan.io and several journal articles [6, 11] would be utilized respectively.

5.1.2 Data Pre-processing and Model Training

Furthermore, the pre-processing is essential for model training as it would affect the accuracy and precision of the classifier model. The dataset would be processed through data cleansing, feature extraction, and attributes integration. Invalid account labels would be filtered based on several heuristics and research conducted in this stages. A variety of machine learning algorithms would be analyzed, compared, and used in training, validating, and testing of the classifier model to optimize the effectiveness in differentiation of the malicious accounts.

5.2 Transaction Analysis with Graphical Visualization

In collection of blockchain data, it is generally conducted through (i) blockchain explorer APIs and (ii) an Ethereum Client Node. To improve the effectiveness of the development, blockchain explorer APIs would be adapted in obtaining specific account data from input (which would be explained in Section 5.3). The usage of API allows more effective request of data with limited resources in spite of there may have restriction for number of transactions obtained from an account.

The account would be analyzed by collecting its details of all transactions. For presenting the transactions graphically, visualization library would be utilized. In

addition, the application may make use of the heuristics-based approaches for clustering of similar transactions into certain groups.

5.3 Application and Front-End Development

In the final stage, the classifier model would be integrated into the web application. As in Section 5.2, the approaches for transaction analysis and enquiry of blockchain data would be embedded into the front-end application. The application would allow user to input the Ethereum account address for analysis. The address could be used in API from Section 5.1, which request and acquire transaction data with the input address. The transaction information including amount of balance, frequency and time of various transaction would present in the UI of the application. The specific address data would be then passed to the classification models in recognizing risks (presence of illicit activities) of the account or transaction. The risks are expected to be displayed in several levels with respect to the result in classification.

In overall, the front-end would be developed in use of Ionic framework with Angular, which is proposed to deploy on Google Firebase. While the ionic framework could provide both web and mobile (IOS, Android) applications, it allows deliverables of cross-platform application. The trained model could be used for classifying the input address data with use of Tensorflow.js and other machine learning tools. In visualization of account transaction, it is proposed for using Sigma.js to present the graphs. Due to undetermined technical challenges in development, the development tools may be subjected to change.

6 Timeline and Milestones

Below is the proposed schedule for various stage of the project:

Stage/ Period	Deadlines	Deliverables and Details
Preliminary Research	Aug – Sep 2023	- Research and doing project plan
Phrase 1 milestones Aug – Sep	**1st Oct 2023**	- Complete Project Plan - Startup of Project Webpage
Data Preparation Oct	**13th Oct 2023**	- Collect blockchain dataset with normal and malicious activity
Illicit Account Classifier Oct – Dec	**1st Dec 2023**	- Algorithms Research - Model Training and Validate - Comparison of Machine learning algorithms
Phrase 2 milestones Oct - Jan	**8th - 12th Jan 2024** (1st presentation) **21st Jan 2024** (Interim Report)	**Interim Report & First Presentation** - Preparation and writing for report and presentation - Preliminary Implementation
Data Analysis with Graphical Visualization Dec - Mar	**1st Mar 2024**	- Organize account data with graphical method - Analyze account data for visualization
Front-End Development & Integration Jan – Mar	**11th Mar 2024**	- Development of application and Integration of classifier model - Allow user input and graphic representations of transaction
Phrase 3 milestones Mar - Apr	**15-19th Apr 2024** (Final presentation) **23rd Apr 2024** (Final report)	- Preparation of Final Presentation and Report - Finalized tested Implementation
Project Exhibition Apr	**26th Apr 2024**	Project Exhibition: - 3-min Video and Poster

7 References

[1] U. Hacioglu, *Blockchain Economics and Financial Market Innovation: Financial Innovations in the Digital Age*. Springer Nature, 2019.

[2] R. M. Aziz, M. F. Baluch, S. Patel, and P. Kumar, "A Machine Learning based Approach to Detect the Ethereum Fraud Transactions with Limited Attributes," *Karbala International Journal of Modern Science*, vol. 8, no. 2, pp. 139–151, May 2022, doi: 10.33640/2405-609x.3229.

[3] Chainalysis, "The 2023 Crypto Crime Report," Chainalysis, Feb. 2023. Accessed: Sep. 21, 2023. [Online]. Available: https://go.chainalysis.com/rs/503-FAP-074/images/Crypto_Crime_Report_2023.pdf

[4] S. Dyson, W. J. Buchanan, and L. Bell, "The challenges of investigating cryptocurrencies and blockchain related crime," *The Journal of the British Blockchain Association*, vol. 1, no. 2, pp. 1–6, Dec. 2018, doi: 10.31585/jbba-1-2-(8)2018.

[5] C. Wang *et al.*, "Demystifying Ethereum account diversity: observations, models and analysis," *Frontiers of Computer Science*, vol. 16, no. 4, Dec. 2021, doi: 10.1007/s11704-021-0221-3.

[6] S. Farrugia, J. Ellul, and G. Azzopardi, "Detection of illicit accounts over the Ethereum blockchain," *Expert Systems With Applications*, vol. 150, p. 113318, Jul. 2020, doi: 10.1016/j.eswa.2020.113318.

[7] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2016.

[8] J. Wu *et al.*, "Who are the phishers? Phishing scam detection on Ethereum via network embedding," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 52, no. 2, pp. 1156–1166, Feb. 2022, doi: 10.1109/tsmc.2020.3016821.

[9] M. Weber *et al.*, "Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics," *arXiv (Cornell University)*, Jul. 2019, doi: 10.48550/arxiv.1908.02591.

[10] H. Han, R. Wang, Y. Chen, K. Xie, and K. Zhang. Research on abnormal transaction detection method for blockchain. In International Conference on Blockchain and Trustworthy Systems, pages 223-236. Springer, 2022.

[11] S. Al-Emari, M. Anbar, Y. K. Sanjalawe, and S. Manickam, "A labeled Transactions-Based dataset on the Ethereum network," in *Springer eBooks*, 2021, pp. 61–79. doi: 10.1007/978-981-33-6835-4_5.

[12] R. M. Aziz, M. F. Baluch, S. Patel, and P. Kumar, "A Machine Learning based Approach to Detect the Ethereum Fraud Transactions with Limited Attributes," *Karbala International Journal of Modern Science*, vol. 8, no. 2, pp. 139–151, May 2022, doi: 10.33640/2405-609x.3229.

[13] P. Li, Y. Xie, X. Xu, J. Zhou, and Q. Xuan, "Phishing fraud detection on Ethereum using Graph neural network," in *Communications in computer and information science*, 2022, pp. 362–375. doi: 10.1007/978-981-19-8043-5_26.

[14] J. Liu *et al.*, "Graph Embedding-Based money laundering detection for Ethereum," *Electronics*, vol. 12, no. 14, p. 3180, Jul. 2023, doi: 10.3390/electronics12143180.

[15] GaoBingyu *et al.*, "Tracking counterfeit cryptocurrency end-to-end," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 4, no. 3, pp. 1–28, Nov. 2020, doi: 10.1145/3428335.